

Amendments to the Claims:

This listing of claims replaces all prior versions and listings of claims in the application:

Listing of Claims:

Claims 1-54 (Canceled).

55. A computer-implemented method for identifying compounds in text, comprising:

extracting a vocabulary of tokens from text;

iterating from $n > 2$ down to $n = 2$ where n decreases by one each iteration and in each iteration performing the actions of:

identifying a plurality of unique n -grams in the text, each n -gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary;

dividing each n -gram into $n-1$ pairs of two adjacent segments, where each segment consists of at least one token;

for each n -gram, calculating a likelihood of collocation for each pair of segments of the n -gram and determining a score for the n -gram based on a lowest calculated likelihood of collocation;

identifying a set of n -grams having scores above a threshold; and

adding the identified set of n -grams as compound tokens to the vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary.

56. The method of claim 55 where calculating a likelihood of collocation for each pair of segments of the n -gram comprises determining a likelihood ratio λ for

each pair of segments that is computed in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of segments.

57. The method of claim 56 where the $L(H_c)$ is computed for each pair of segments, t_1, t_2 , in each n -gram in accordance with the formula:

$$\arg \max_{L(H_i)} \frac{L(t_1, t_2 \text{ form compound})}{L(n - \text{gram does not form compound})}.$$

58. The method of claim 56 where, for each pair of segments, t_1, t_2 , in each n -gram, the independence hypothesis comprises $P(t_2 | t_1) = P(t_2 | \bar{t}_1)$ and the collocation hypothesis comprises $P(t_2 | t_1) > P(t_2 | \bar{t}_1)$.

59. The method of claim 55 where identifying a plurality of unique n -grams in the text comprises skipping n -grams appearing in a list of known compounds.

60. A computer program product, encoded on a computer-readable medium, operable to cause data processing apparatus to perform operations comprising:

extracting a vocabulary of tokens from text;

iterating from $n > 2$ down to $n = 2$ where n decreases by one each iteration and in each iteration performing the actions of:

identifying a plurality of unique n -grams in the text, each n -gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary;

dividing each n -gram into $n-1$ pairs of two adjacent segments, where each segment consists of at least one token;

for each n-gram, calculating a likelihood of collocation for each pair of segments of the n-gram and determining a score for the n-gram based on a lowest calculated likelihood of collocation;

identifying a set of n-grams having scores above a threshold; and
 adding the identified set of n-grams as compound tokens to the vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary.

61. The program product of claim 60 where calculating a likelihood of collocation for each pair of segments of the n-gram comprises determining a likelihood ratio λ for each pair of segments that is computed in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of segments.

62. The program product of claim 61 where the $L(H_c)$ is computed for each pair of segments, t_1, t_2 , in each n-gram in accordance with the formula:

$$\arg \max_{L(H_i)} \frac{L(t_1, t_2 \text{ form compound})}{L(n - \text{gram does not form compound})}.$$

63. The program product of claim 61 where, for each pair of segments, t_1, t_2 , in each n-gram, the independence hypothesis comprises $P(t_2 | t_1) = P(t_2 | \bar{t}_1)$ and the collocation hypothesis comprises $P(t_2 | t_1) > P(t_2 | \bar{t}_1)$.

64. The program product of claim 60 where identifying a plurality of unique n -grams in the text comprises skipping n -grams appearing in a list of known compounds.

65. A system comprising:

- a computer readable medium including a program product; and
- one or more processors configured to execute the program product and perform operations comprising:

- extracting a vocabulary of tokens from text;

- iterating from $n > 2$ down to $n = 2$ where n decreases by one each iteration and in each iteration performing the actions of:

- identifying a plurality of unique n -grams in the text, each n -gram being an occurrence in the text of n sequential tokens, each token being found in the vocabulary;

- dividing each n -gram into $n-1$ pairs of two adjacent segments, where each segment consists of at least one token;

- for each n -gram, calculating a likelihood of collocation for each pair of segments of the n -gram and determining a score for the n -gram based on a lowest calculated likelihood of collocation;

- identifying a set of n -grams having scores above a threshold; and
 - adding the identified set of n -grams as compound tokens to the vocabulary and removing constituent tokens that occur in the added compound tokens from the vocabulary.

66. The system of claim 65 where calculating a likelihood of collocation for each pair of segments of the n -gram comprises determining a likelihood ratio λ for each pair of segments that is computed in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis,

$L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of segments.

67. The system of claim 66 where the $L(H_c)$ is computed for each pair of segments, t_1, t_2 , in each n -gram in accordance with the formula:

$$\operatorname{argmax}_{L(H_c)} \frac{L(t_1, t_2 \text{ form compound})}{L(n\text{-gram does not form compound})}.$$

68. The system of claim 66 where, for each pair of segments, t_1, t_2 , in each n -gram, the independence hypothesis comprises $P(t_2 | t_1) = P(t_2 | \bar{t}_1)$ and the collocation hypothesis comprises $P(t_2 | t_1) > P(t_2 | \bar{t}_1)$.

69. The system of claim 65 where identifying a plurality of unique n -grams in the text comprises skipping n -grams appearing in a list of known compounds.